# Personal Privacy Vs. Information Age

Wenhui Fang
GR 5243: Advanced Data Science
Columbia University, Department of Statistics

# Introduction:

- Personal privacy vs. Information Age, a hot topic over many years.
- Increasing number of stalkers using social media information.
- Weibo began to display user’s IP address without user’s consent
  - Discussion about personal privacy invaded
  - Discussion about bad people can use this to do

# Project Goal:

- How easily personal information can be obtained?
- How our information can be used in good or bad ways?

# Data Collection

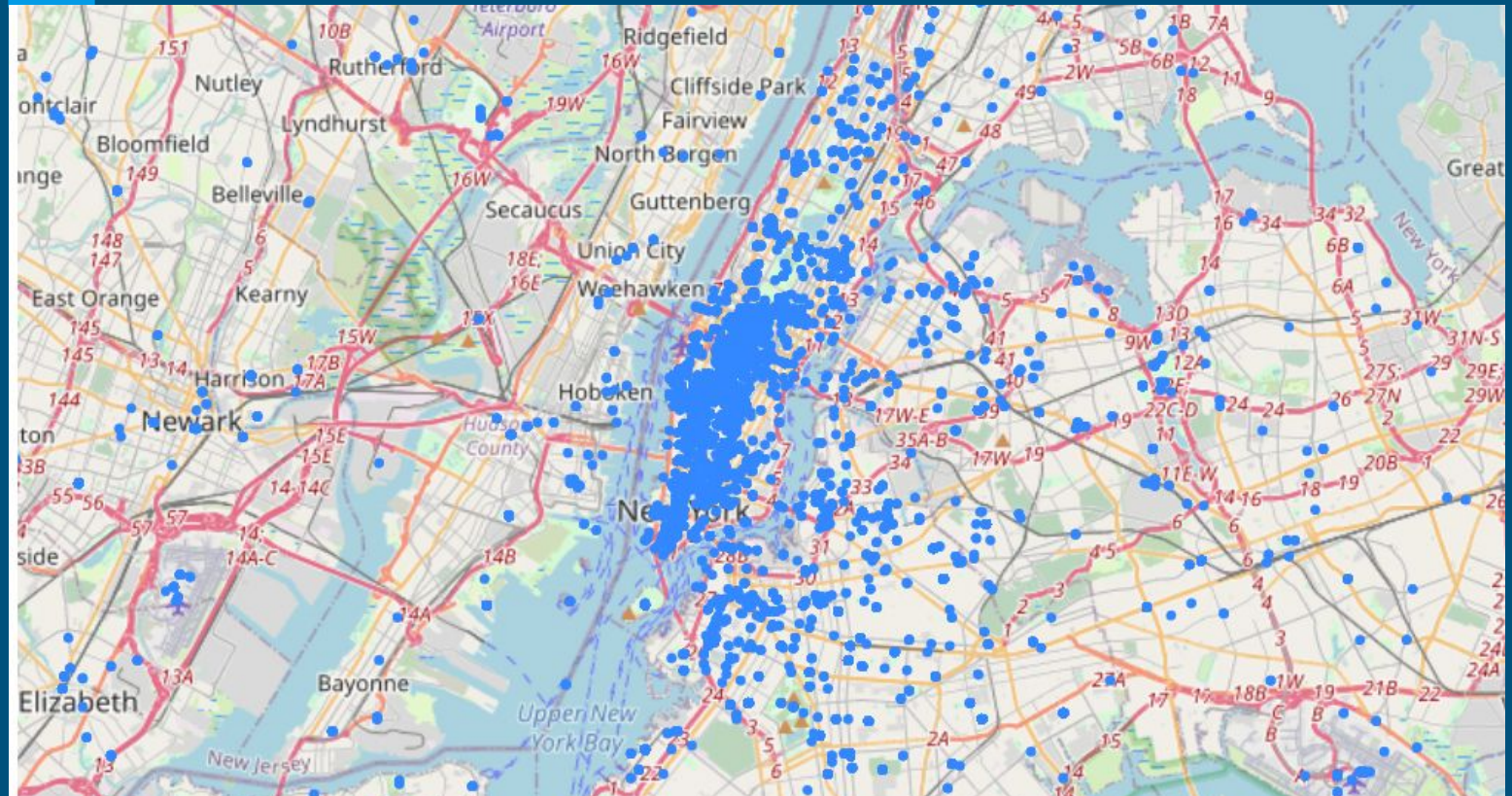
- Some Tweets comes with geographic location.
- If we gather enough tweet with different location, we can infer on one's travel pattern.
- Use snscrapper package in python to find geotagged tweets, save usernames.
- Input usernames to Tweepy package for full tweets history for each user
- Dataset consists of 65170 tweets.
- After cleaning, we have 27126 tweets and 591 users with their locations, and time for each tweet.

# Data Overview

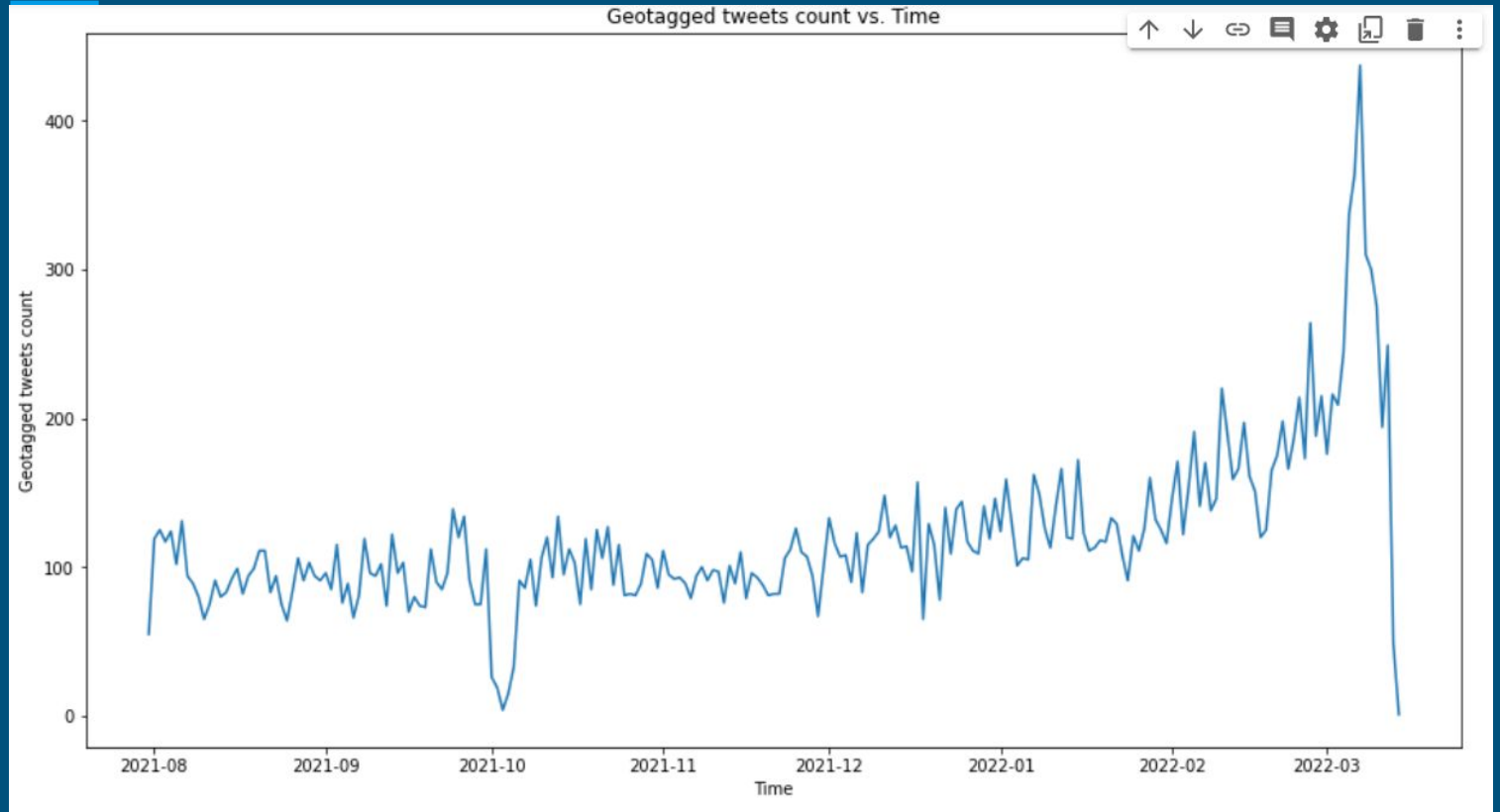
- Every tweet comes with latitude and longitude that have precision with 4 digits.
- Time of tweet also specified to second.

| | date | year | month | day | hour | minute | second | user_id | user | geo | latitude | longitude |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2021-12-27 | 2021 | 12 | 27 | 11 | 53 | 59 | 24402703.0 | MerDiann | 41.2225, -74.2897 | 41.2225 | -74.2897 |
| **1** | 2021-11-30 | 2021 | 11 | 30 | 18 | 26 | 49 | 24402703.0 | MerDiann | 35.9886, -78.9072 | 35.9886 | -78.9072 |

# Visualization



# Visualization



# Set-up

- Divided longitude and latitude equally to create different zones.
  - S = {z1, z2, ..., zn}
- Divide time into 24 intervals by one hour.
  - T = {t1, t2, ..., tn}
- By calculating the ratio of traveler's record falls into zone Z and time T, we have the  MLE of probability showing up in Z during T
- Let S be column and T be rows: form a probability distribution matrix for each traveler.
  - Denoted by S-T matrix

# Features using in Predict Travel Mode

- Spatial Distribution Similarity (AZ): (Travel space similarity)
  - Each traveler's spatial distribution vector is obtained from S-T matrix, and the similarity between two travelers is calculated by cosine similarity.
- Temporal Distribution Similarity (AT): (Travel time similarity)
  - Each traveler's temporal distribution vector obtain is from S-T matrix, and the similarity is also calculated by cosine similarity
- Radius of gyration (AR): (Travel radius)
  - Gyration is calculated by the standard deviation of a travel's spatial distance. Similarity of gyration is calculated by
  - $cl(x,y)=1-2\times|sigmoid(x-y)-0.5|$
- Travel frequency similarity:
  - Frequency is calculated by the ratio of number of travel records and the number of observation period. Similarity of frequency is also calculated by cl(x,y) function

# Reference Label

- The reference label is overall travel pattern similarity
- $$S_{ij} = \sum_{m=1}^{T} \sum_{k=1}^{N} \sqrt{p_{mk}^{i} * p_{mk}^{i}},$$
- Where T is the number of time span, N is the number of zones.
- p is MLE probability in ST matrix

# Model: Multivariate Linear Regression

- Since we have four features that are correlated with reference label, we use multivariate linear regression.
- We got R square = 0.806
  - 80% variation in overall travel pattern similarity has been explained by 4 features
- All p-values for features are less than 0.05
  - Preserve them all in the model

# Conclusion:

- Relatively easy to obtain user’s personal information
- After performing similarity analysis, a relatively simple model can have a high accuracy predicting one’s travel pattern
- Dataset are not just digits, each case is living life
- We can use this analysis in many ways such as travel, transportation
- It can also be used to predict one specific person’s travel mode

# Conclusion:

- In this information age, big data provides data scientists a way to find solutions efficiently.
- Personal information that are not protected well can be used in bad way

# Thank you!